

# Package ‘wordpiece.data’

October 8, 2021

**Title** Data for Wordpiece-Style Tokenization

**Version** 1.0.2

**Description** Provides data to be used by the wordpiece algorithm in order to tokenize text into somewhat meaningful chunks. Included vocabularies were retrieved from <https://huggingface.co/bert-base-cased/resolve/main/vocab.txt> and <https://huggingface.co/bert-base-uncased/resolve/main/vocab.txt> and parsed into an R-friendly format.

**License** Apache License (>= 2)

**Encoding** UTF-8

**RoxygenNote** 7.1.2

**URL** <https://github.com/macmillancontentscience/wordpiece.data>

**BugReports** <https://github.com/macmillancontentscience/wordpiece.data/issues>

**Depends** R (>= 3.5.0)

**Suggests** testthat (>= 3.0.0)

**Config/testthat/edition** 3

**NeedsCompilation** no

**Author** Jonathan Bratt [aut] (<<https://orcid.org/0000-0003-2859-0076>>),  
Jon Harmon [aut, cre] (<<https://orcid.org/0000-0003-4781-4346>>),  
Bedford Freeman & Worth Pub Grp LLC DBA Macmillan Learning [cph],  
Google, Inc [cph] (original BERT vocabularies)

**Maintainer** Jon Harmon <[jonthegeek@gmail.com](mailto:jonthegeek@gmail.com)>

**Repository** CRAN

**Date/Publication** 2021-10-08 21:50:02 UTC

## R topics documented:

|                           |          |
|---------------------------|----------|
| wordpiece_vocab . . . . . | 2        |
| <b>Index</b>              | <b>3</b> |

---

|                 |                                    |
|-----------------|------------------------------------|
| wordpiece_vocab | <i>Load a wordpiece Vocabulary</i> |
|-----------------|------------------------------------|

---

**Description**

A wordpiece vocabulary is a named integer vector with class "wordpiece\_vocabulary". The names of the vector are the tokens, and the values are the integer identifiers of those tokens. The vocabulary is 0-indexed for compatibility with Python implementations.

**Usage**

```
wordpiece_vocab(cased = FALSE)
```

**Arguments**

`cased` Logical; load the uncased vocabulary, or the cased vocabulary?

**Value**

A wordpiece\_vocabulary.

**Examples**

```
head(wordpiece_vocab())  
head(wordpiece_vocab(cased = TRUE))
```

# Index

wordpiece\_vocab, [2](#)